406 final proj final

qwluo

2023-12-08

Data

One of our datasets, titled "Global YouTube Statistics," offers a comprehensive overview of the top YouTube channels, encompassing a wide range of metrics such as subscriber counts, video views, upload frequency, and, crucially, earnings data. It provides detailed information on the performance of popular Youtube channels across multiple countries and content categories. Kay variables such as highest yearly earnings allow us to quantify video success and revenue generated on the platform.

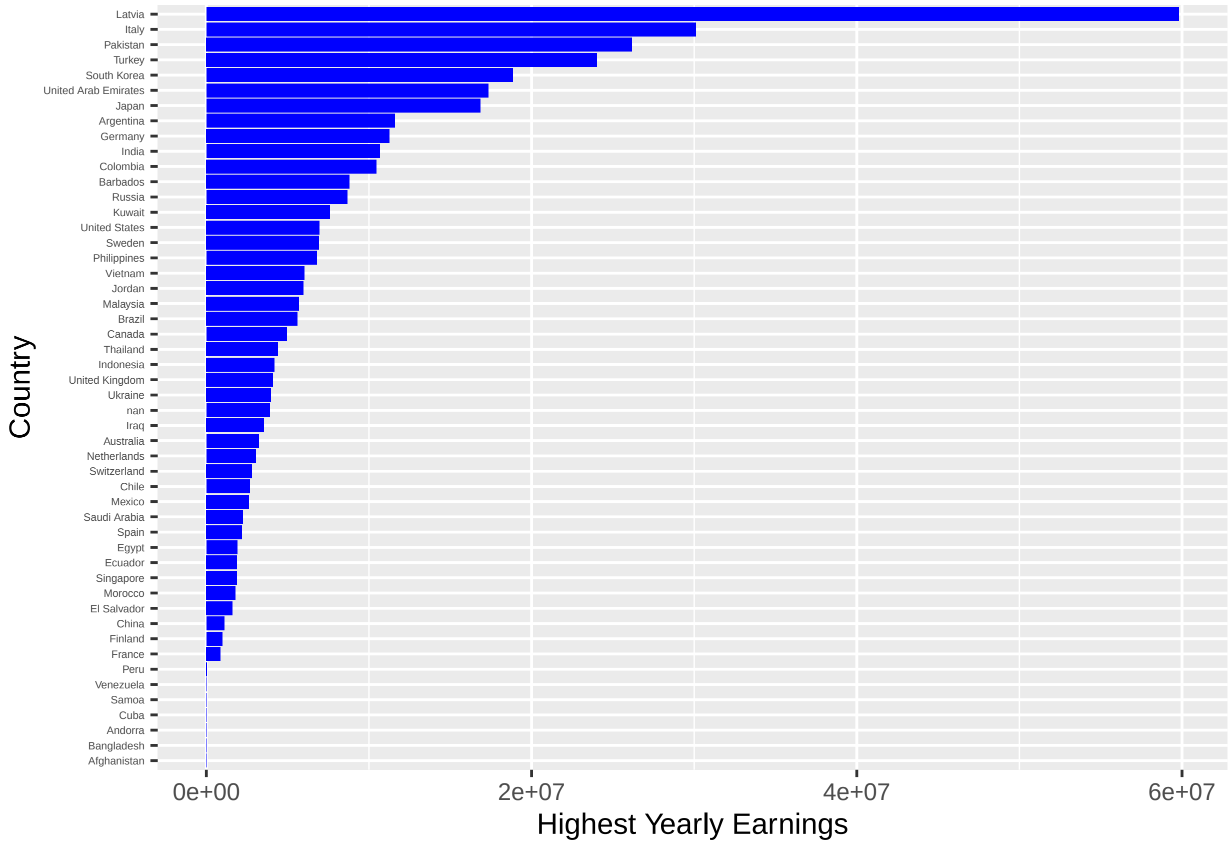
The country and category associated with each channel enables comparative analysis across these groupings. The country and category associated with each channel enables comparative analysis across these groupings. For example, we can analyze whether significant differences exist in earnings between channels from the United States versus India. Or between music channels versus comedy channels.

The ANOVA method is well-suited for this comparison as it statistically tests if the mean earnings differs between various countries and categories. A significant ANOVA F-test suggests location and content-type impacts revenue potential on YouTube.

We can further validate these findings using permutation testing. By randomly shuffling the country and category labels, we can simulate the null hypothesis of no true differences between groups. Comparing the real observation to distributions under this null hypothesis provides a non-parametric test of significance.

To achieve a nuanced understanding of these aspects, advanced statistical techniques, including ANOVA (Analysis of Variance) and permutation tests, were employed. These methods provided insights into whether the differences observed in earnings based on country and category are statistically significant or merely products of random variation.

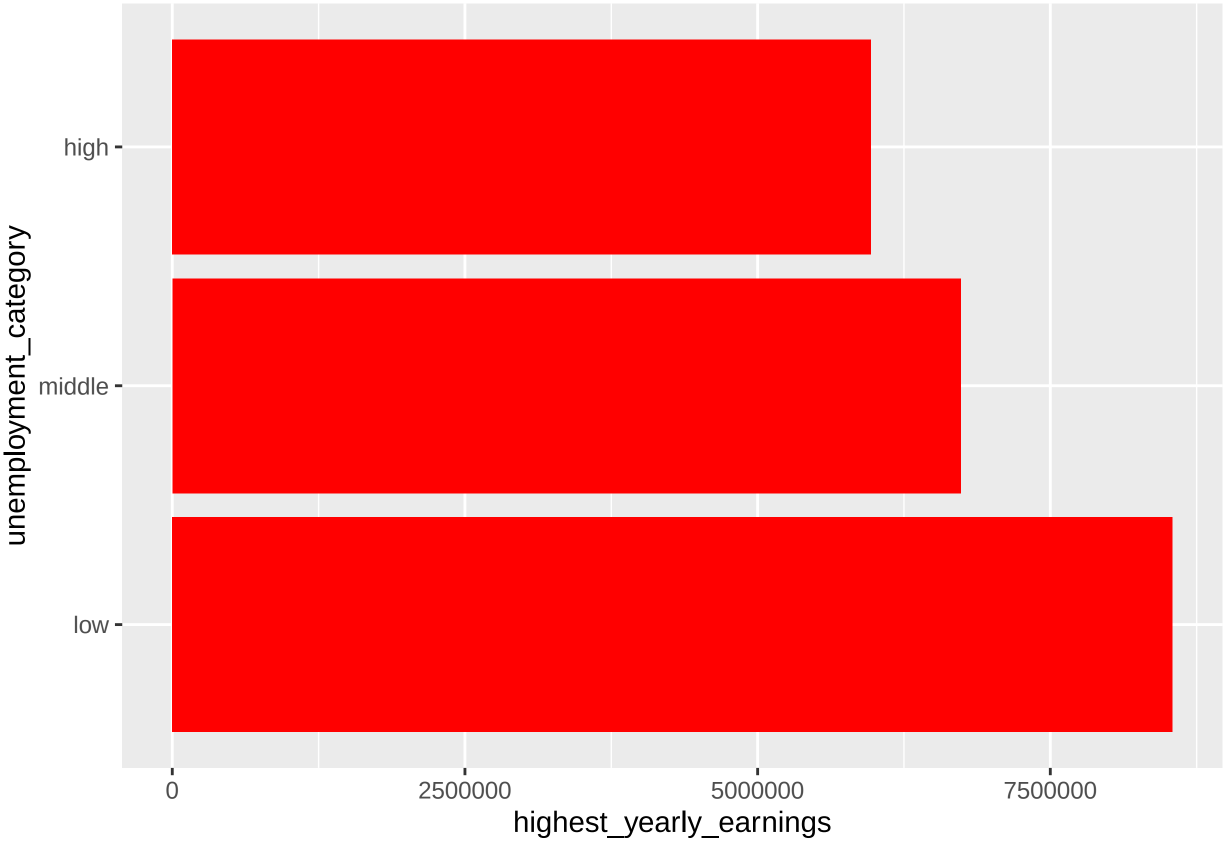
Data Processing and Cleaning



There's a noticeable disparity in the average highest yearly earnings among YouTube channels from different countries, ranging from over $60,000 for channels based in Latvia down to almost zero earnings for some lower-income countries. This suggests that the location of a channel can significantly impact its earning potential. The top-earning countries might have a combination of factors such as larger audience bases, higher advertising rates, and more lucrative sponsorship opportunities.

Specifically, the highest-earning countries tend to be more developed nations such as Latvia and Italy potentially due to larger audience bases, higher advertising rates, and more lucrative sponsorship opportunities. For example, Latvia's highly educated population, high level of urbanization with widespread internet access enabling content distribution, and integration with European advertising markets likely assist Latvian YouTubers in commanding top earnings on the platform globally.

Conversely, lower-income developing countries face structural disadvantages in maximizing the profitability of YouTube channels. Afghanistan, Bangladesh, Andorra, Cuba, Samoa, and Venezuela registered close to zero average earnings, as indicated by the histogram. Limited internet connectivity, lower viewership demand, and lack of prominent advertiser interest make it difficult for creators in these geographies to monetize content. Add on political and economic instability, prevalent in countries like Venezuela, leading to even sharper drops in viable revenue sources.



Different level of unemployment categories on YouTube have varying average highest yearly earnings. The bar representing the low unemployment category has the highest average yearly earnings, exceeding 7,500,000. This suggests that in areas with low unemployment rates, the highest yearly earnings tend to be greater. This could be indicative of stronger economic conditions in these areas, potentially leading to higher disposable incomes and greater spending power. The high unemployment category has the lowest average highest yearly earnings, around 6,000,000. This suggests that areas with higher unemployment rates tend to have lower highest yearly earnings.

The observed trend suggests a correlation between unemployment levels and economic success, as measured by highest yearly earnings. Lower unemployment rates might be associated with more robust economic activity, leading to higher earnings. It highlights the influence of broader socio-economic factors on the commercial success of content creators, suggesting a nuanced relationship between employment status and digital engagement.

Method

This report employs two primary statistical methods to analyze the dataset: Analysis of Variance (ANOVA) and Permutation Testing.

1. **Analysis of Variance (ANOVA)**

   ANOVA is used to compare the means of three or more independent groups to determine if there is any statistically significant difference between them. The formula for the one-way ANOVA F-test statistic is:

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

where MS_{between} (Mean Square Between Groups) and MS_{within} (Mean Square Within Groups) are given by:

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}}$$
$$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}}$$

Here, SS_{between} (Sum of Squares Between Groups) and SS_{within} (Sum of Squares Within Groups) are calculated as follows:

$$SS_{\text{between}} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$$
$$SS_{\text{within}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$$

where k is the number of groups, n_i is the number of observations in group i, $\bar{X}_i$ is the mean of group i, X_{ij} is the jth observation in group i, and $\bar{X}$ is the overall mean.

In this analysis, ANOVA was applied to compare the success index across different countries and urbanization categories.

2. **Permutation Testing**

   The permutation test is a non-parametric method used to determine the significance of the difference between two groups. This test involves repeatedly shuffling the data and recalculating the test statistic to create a distribution of the statistic under the null hypothesis.

   The test statistic used is the difference in means between the two groups. The process is as follows:

   - Combine the data from both groups into a single dataset.
   - For a large number of permutations (e.g., 1000):
     - Randomly shuffle the combined dataset.
     - Split the shuffled dataset into two new groups, maintaining the original group sizes.
     - Calculate the difference in means between these two new groups.
   - Calculate the proportion of permutations where the difference in means is greater than the observed difference. This proportion is the p-value.

ANOVA and Analyse

The ANOVA test has been conducted to compare the highest yearly earnings of YouTube channels based on their country of origin. The results are as follows:

```
##               Df    Sum Sq   Mean Sq F value   Pr(>F)    
## Country       49 1.825e+16 3.724e+14   2.058 3.87e-05 ***
## Residuals    945 1.710e+17 1.809e+14                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-Statistic is a measure of the variance between the group means relative to the variance within the groups. The F-statistic of 2.058 compares the variance between the group means (between different countries) to the variance within the groups (within each country). A higher value indicates a greater difference between the groups so this is a strong indication that the country of origin plays a notable role in influencing the earnings of each video.

The P-value is particularly important. In this case, the P-value is approximately 0.0000387, which is less than the commonly used significance level of 0.05. This suggests that there are statistically significant differences in the highest yearly earnings among YouTube channels from different countries. In other words, the country of origin appears to have a significant impact on the highest yearly earnings of YouTube channels in this dataset. Given the very small p-value, we can reject the null hypothesis that there is no difference in the mean highest earnings across countries.

The ANOVA test has been conducted to compare the highest yearly earnings of YouTube channels across different unemployment categories. The results are as follows:

```
##                        Df    Sum Sq   Mean Sq F value Pr(>F)
## unemployment_category   2 7.091e+14 3.546e+14   1.811  0.164
## Residuals             869 1.701e+17 1.958e+14
```

The test has two components with different degrees of freedom (Df). For the 'unemployment_category', Df is 2, indicating three unemployment categories are being compared (low, medium, and high). The F-value of 1.811 suggests the ratio of the variance explained by the unemployment categories to the variance within the categories (unexplained variance). An F-value greater than 1 suggests that there is more variability between the groups than within them. The p-value of 0.164 is above the conventional alpha level of 0.05. This indicates that the differences in the highest yearly earnings among YouTube channels across different unemployment categories are not statistically significant. In other words, based on this data and this test, we do not have sufficient evidence to conclude that the unemployment category has a significant impact on the highest yearly earnings of YouTube channels. Thus, the ANOVA results suggest that the highest yearly earnings of YouTube channels do not differ significantly based on the unemployment category, based on the data analyzed.

Permutation Test Simulation

The permutation test was conducted to evaluate whether there are statistically significant differences in the highest yearly earnings of YouTube channels across different countries.

Methodology Observed Difference Calculation: The observed difference in mean earnings was calculated as the difference between the maximum and minimum mean earnings across all countries. This value represents the largest observed disparity in earnings in the dataset.

Permutation Test Execution: We performed the permutation test 10,000 times. In each iteration, two countries were randomly selected. The data was then subsetted to include only these countries. Within this subset, the highest yearly earnings were shuffled, and the difference in mean earnings (perm_diff[i]) was recalculated for this shuffled subset.

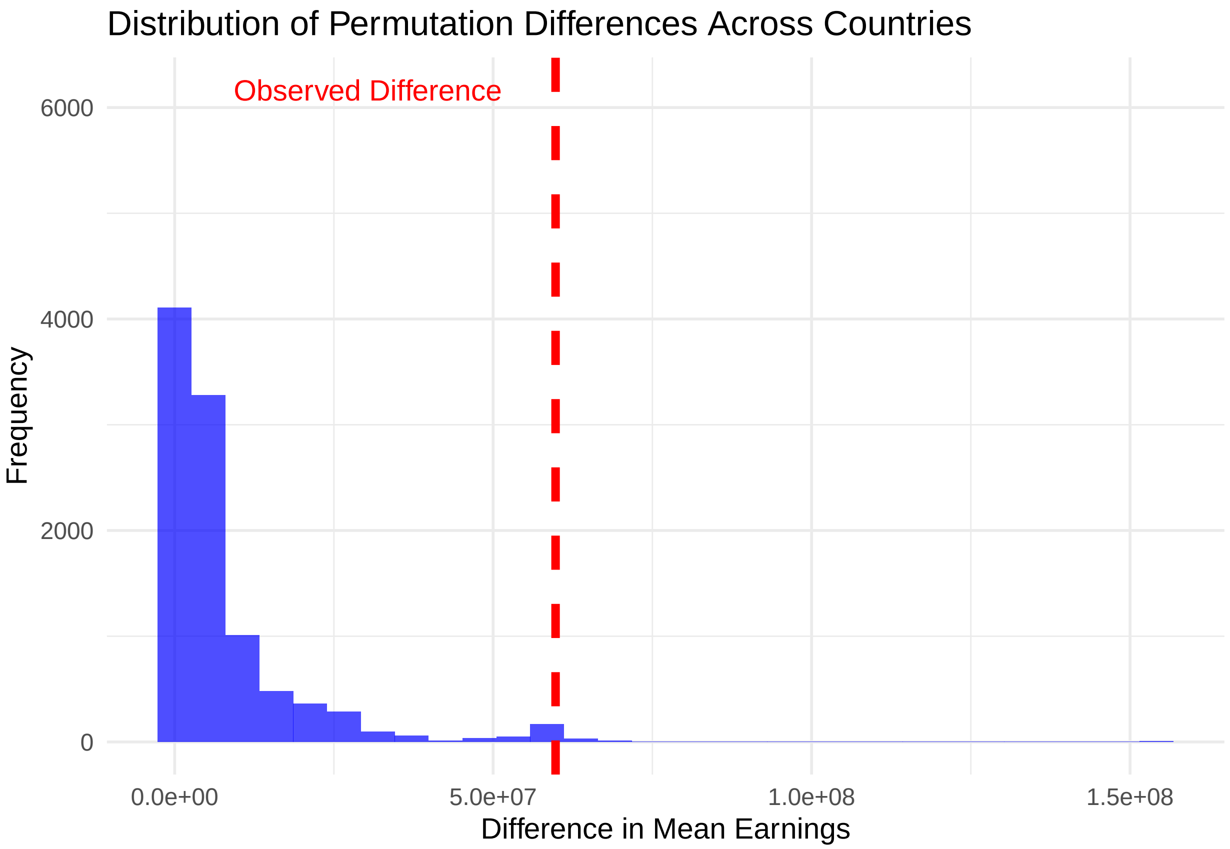
P-Value Calculation: The p-value was determined by comparing the observed difference against the distribution of differences obtained from the permutations. Specifically, it was calculated as the proportion of permutation differences that were greater than or equal to the observed difference.

```
## [1] 0.0068
```

P-Value: The permutation test yielded a p-value of 0.0068. This indicates that only 0.68% of the permutation differences were as large as or larger than the observed difference.

Statistical Significance: Contrary to the initial interpretation, a p-value of 0.0068 is actually below the conventional alpha level of 0.05. This suggests that the observed difference in highest yearly earnings across countries is statistically significant. Null Hypothesis Rejection: Given the low p-value, we reject the null hypothesis. There is statistically significant evidence to suggest that the country of origin does impact the highest yearly earnings of YouTube channels.

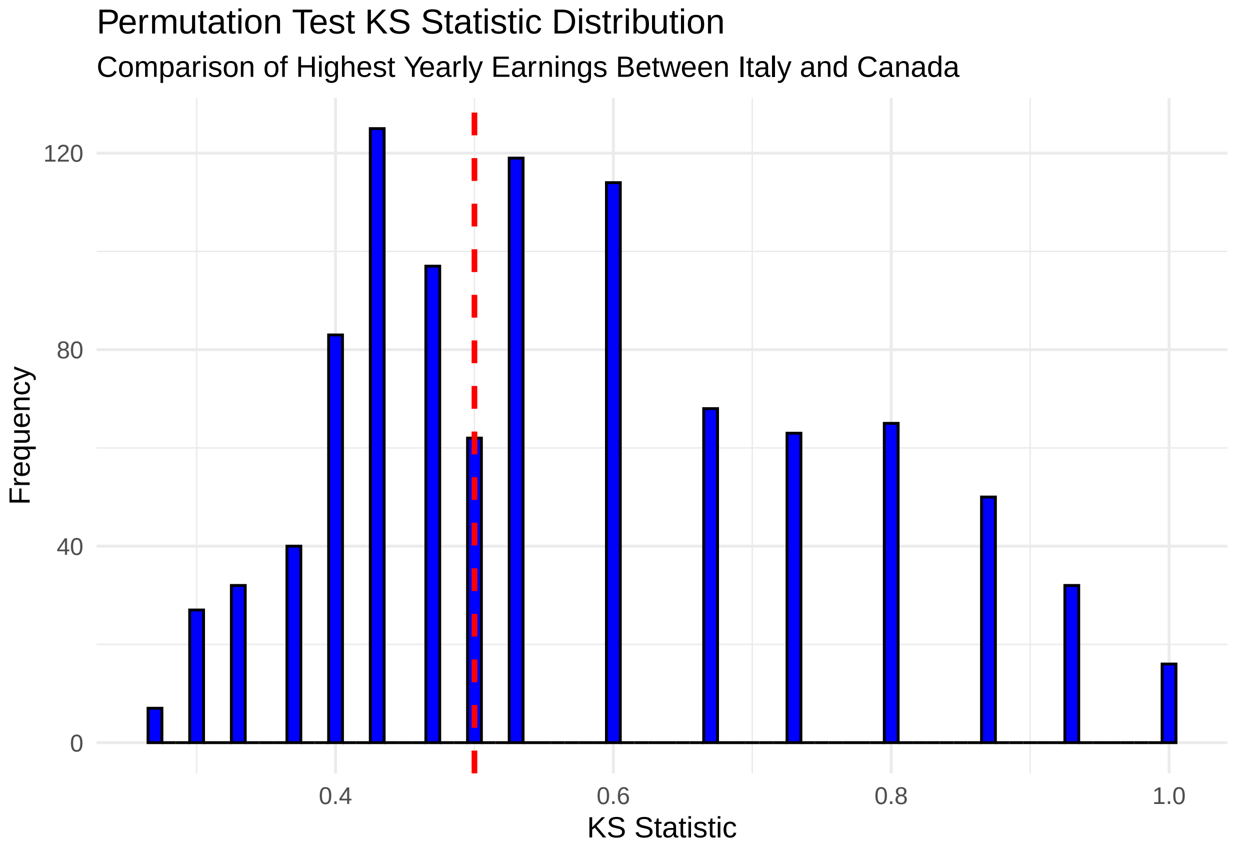
```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



A histogram was created to visualize the distribution of permutation differences. The observed difference was marked with a red dashed line for reference.

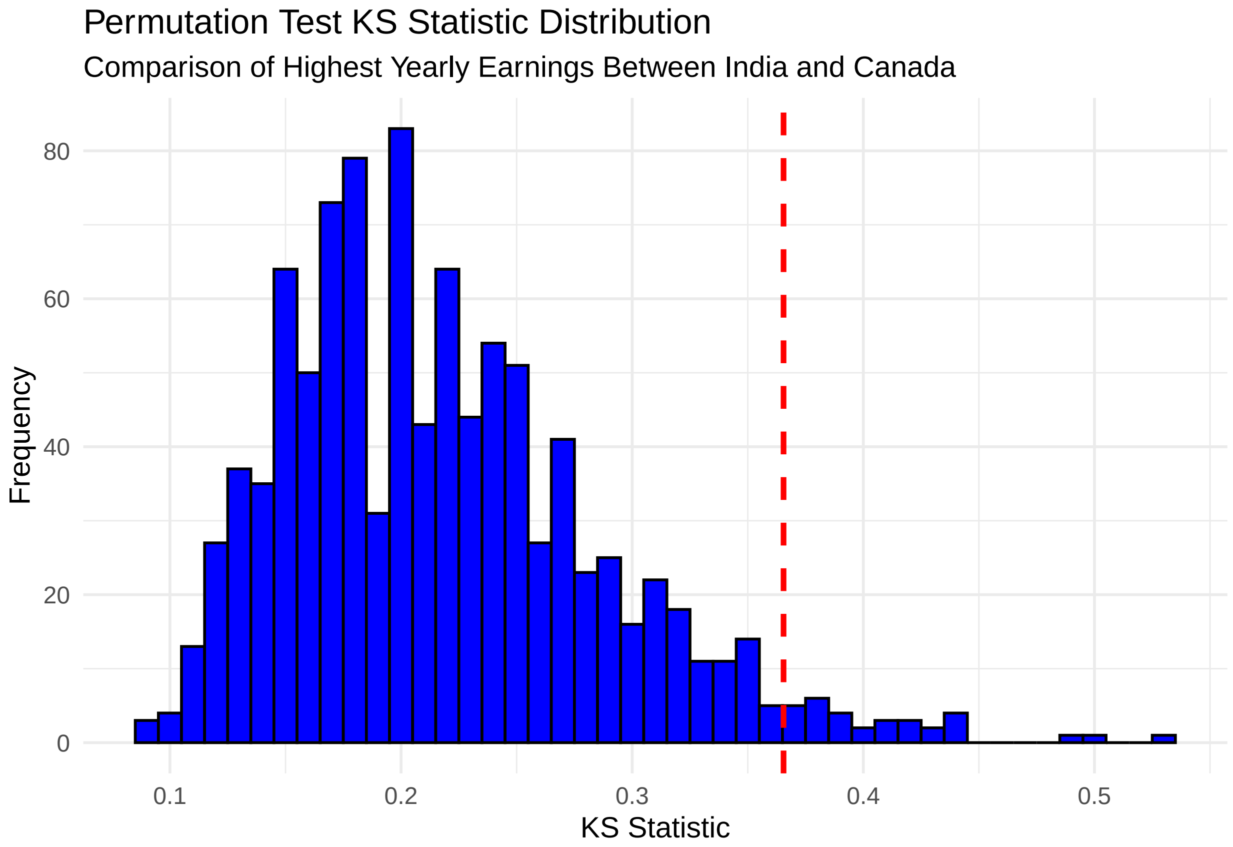
Histogram Interpretation: The histogram shows the range and frequency of differences obtained from the permutations. The placement of the observed difference (red line) towards the extremes of this distribution visually supports the statistical finding that the observed difference is unusual compared to the permutation results.

While a general permutation test across all countries can reveal overall differences in earnings, it doesn't provide insights into how specific countries compare. The KS test for specific pairs allows for a more detailed, pairwise comparison, helping to understand the relationship between two particular countries. The KS test is particularly effective in detecting differences not just in the means, but in the overall distribution of data. This means it can identify variations in the spread, skewness, or presence of outliers in earnings between two countries, offering a more comprehensive view of the differences. The KS test provides a non-parametric method to compare distributions, making it robust against non-normal data, which is often the case in real-world datasets like YouTube earnings.



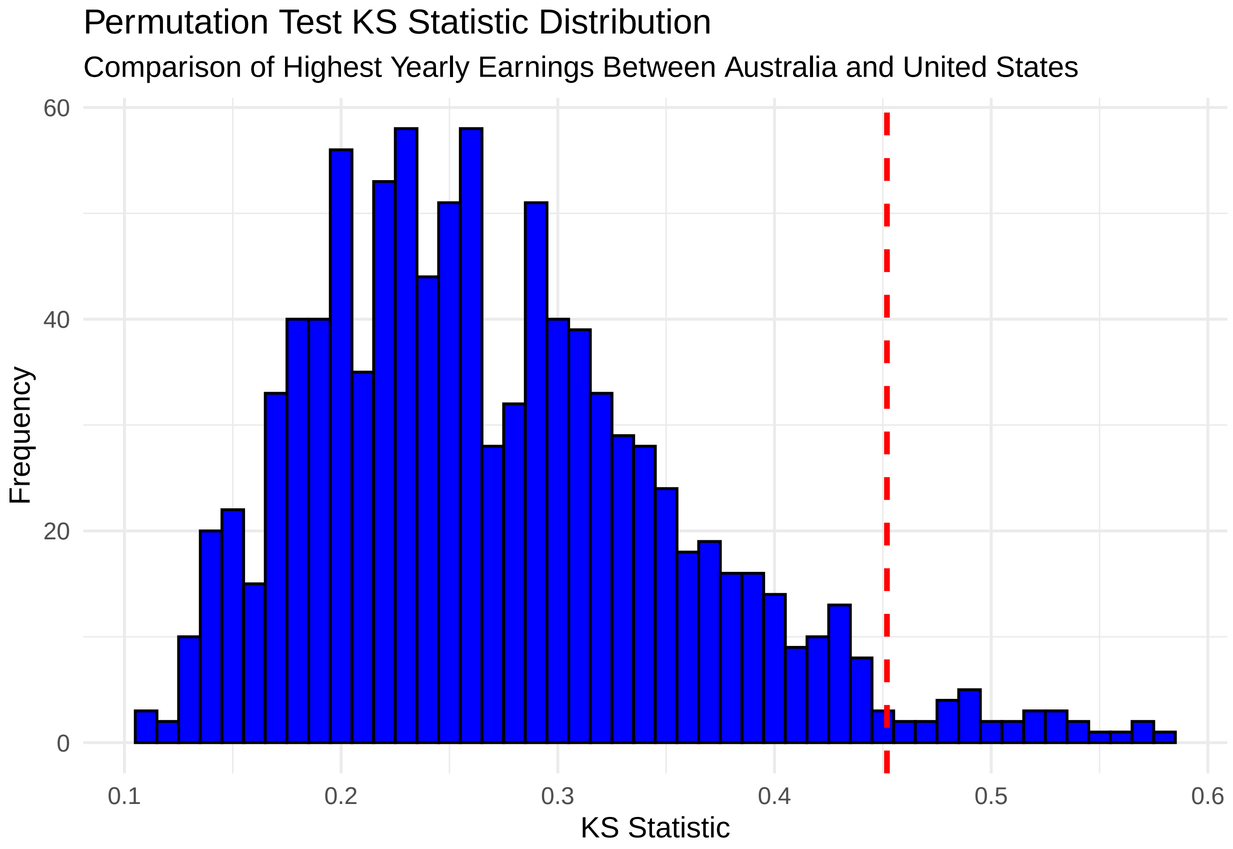
```
## [1] 0.568
```

The high p-value indicates that there is no statistically significant difference in the earnings distributions between YouTube channels in Italy and Canada. This suggests that the earnings profiles for these two countries are quite similar.



```
## [1] 0.031
```

We obtained a p-value of 0.031 here. The low p-value here suggests a statistically significant difference in the earnings distributions between YouTube channels in India and Canada, which implies that the earnings profiles for these countries are distinct from each other. This might reflect differences in market conditions, audience size, or monetization potential in these countries.



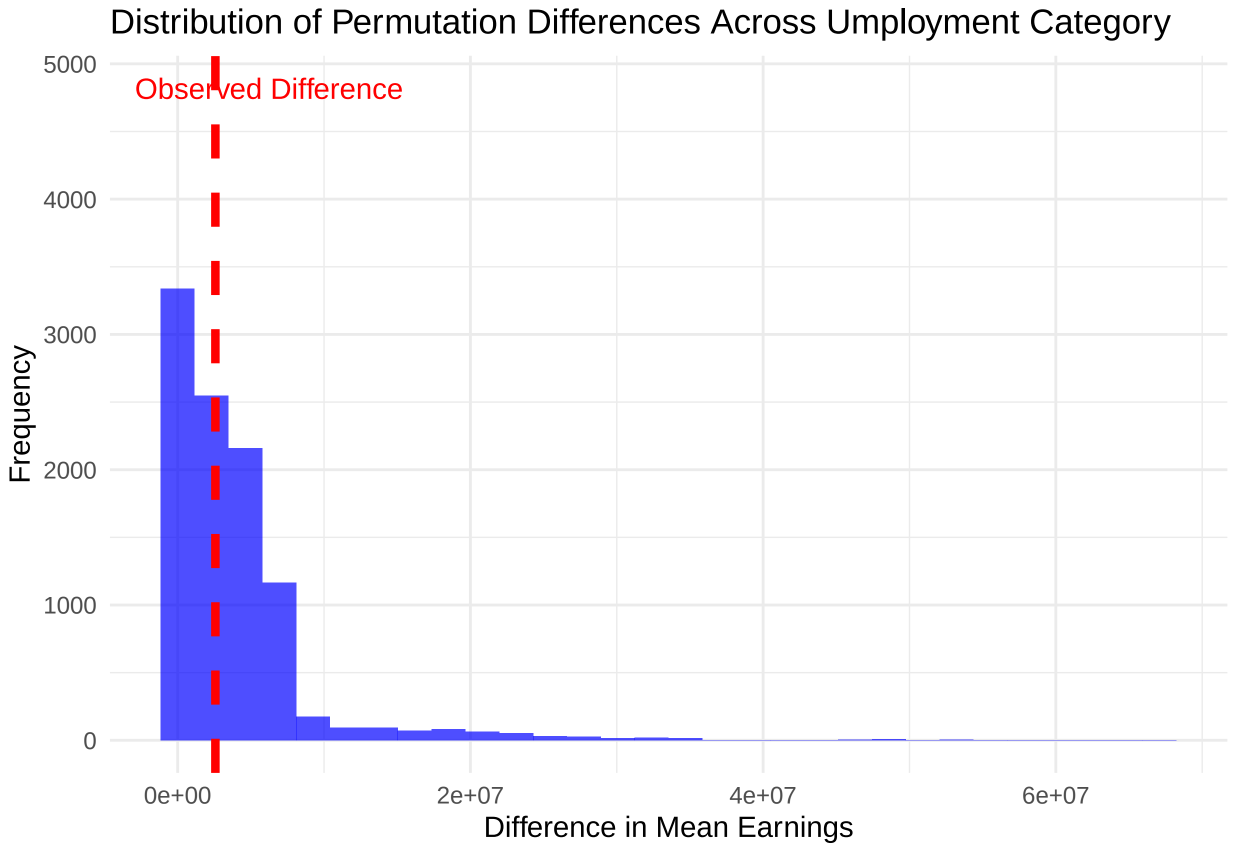
```
## [1] 0.03
```

Similar to the India-Canada comparison, we get a p-value of 0.03. This low p-value indicates a significant difference in the earnings distributions between YouTube channels in Australia and the United States.

Therefore, while the general permutation test shows significant variability in earnings across all countries, the KS tests reveal that this variability is not uniform across all pairs of countries. This suggests that while there is a general global trend, specific country-to-country comparisons can yield different insights. The differences in p-values between specific country pairs (like India-Canada vs. Italy-Canada) suggest that country-specific factors (such as market conditions, audience preferences, economic size, or content types) significantly influence earnings. Some country pairs have similar earnings profiles, while others differ markedly. Also The KS test is sensitive to differences in both the location and shape of the earnings distributions. This means it can detect differences not just in the averages, but also in the overall distribution patterns such as variability and skewness.

we also want to assess whether there are statistically significant differences in the highest yearly earnings of YouTube channels across different levels of unemployment. The p-value obtained from this test was 0.4862. A p-value of 0.4862 is well above the conventional alpha level of 0.05. This indicates that the observed difference in mean earnings across different unemployment categories is not statistically significant. With such a high p-value, you fail to reject the null hypothesis. The null hypothesis in this context posits that there is no significant difference in the highest yearly earnings across different unemployment categories. Based on our data, the level of unemployment does not have a statistically significant impact on the highest yearly earnings of YouTube channels.

```
## [1] 0.4862
```



The result suggests that the category of unemployment, used here as an indicator of broader economic conditions, does not appear to be a primary factor influencing the financial success of YouTube channels. The histogram visualizes the distribution of permutation differences after shuffling the unemployment categories - essentially simulating the null case of no category impact. Importantly, the observed difference, shown by the red line, lies well within this distribution rather than in the tails. Only 48.62% of the permutation differences were greater than the actual observation. This suggests the observed difference could plausibly occur just by chance even if unemployment levels do not truly alter earnings.

Therefore, we conclude that countries' unemployment status does not have a statistically detectable effect on the highest yearly YouTube earnings channels located there. Variability within categories is much larger than the variability between categories.

This outcome suggests that macroeconomic factors, such as employment conditions, may not be directly influential in shaping the revenue potential of individual YouTube channels. Instead, factors specific to the content creator, such as the quality of content, audience engagement, and sponsorship deals, are likely more pivotal in determining financial success. While the geographical location of a channel does play a significant role in its earnings, finer economic indicators like unemployment rates do not seem to dictate success on YouTube. Instead, the financial outcomes for YouTube channels are more likely to be reshaped by factors related to audience interaction and the nature of the content itself, rather than by the broader economic environment, such as employment rates or population size.